Using Minimal Recursion Semantics for Entailment Recognition

Elisabeth Lien
Department of Informatics, University of Oslo
elien@ifi.uio.no

Abstract

This paper describes work on using Minimal Recursion Semantics (MRS) representations for the task of recognising textual entailment. I use entailment data from a SemEval-2010 shared task to develop and evaluate an entailment recognition heuristic. I compare my results to the shared task winner, and discuss differences in approaches. Finally, I run my system with multiple MRS representations per sentence, and show that this improves the recognition results for positive entailment sentence pairs.

1 Introduction

Since the first shared task on Recognising Textual Entailment (RTE) (Dagan et al., 2005) was organised in 2005, much research has been done on how one can detect entailment between natural language sentences. A range of methods within statistical, rule based, and logical approaches have been applied. The methods have exploited knowledge on lexical relations, syntactic and semantic knowledge, and logical representations.

In this paper, I examine the benefits and possible disadvantages of using rich semantic representations as the basis for entailment recognition. More specifically, I use Minimal Recursion Semantics (MRS) (Copestake et al., 2005) representations as output by the English Resource Grammar (ERG) (Flickinger, 2000). I want to investigate how logical-form semantics compares to syntactic analysis on the task of determining the entailment relationship between two sentences. To my knowledge, MRS representations have so far not been extensively used for this task.

To this end, I revisit a SemEval shared task from 2010 that used entailment recognition as a means to evaluate parser output. The shared task data were constructed so as to require only syntactic analysis to decide entailment for a sentence pair. The MRSs should perform well on such data, as they abstract over irrelevant syntactic variation, as for example use of active vs. passive voice, or meaning-preserving variation in constituent order, and thus normalise at a highly suitable level of "who did what to whom". The core idea of my approach is graph alignment over MRS representations, where successful alignment of MRS nodes is treated as an indicator of entailment.

This work is part of an ongoing dissertation project, where the larger goal is to look more closely at correspondences between logical and textual entailment, and the use of semantic representations in entailment recognition.

Besides using MRS, one novel aspect of this work is an investigation of using n-best lists of parser outputs in deciding on entailment relations. In principle, the top-ranked (i.e., most probable) parser output should correspond to the intended reading, but in practise this may not always be the case. To increase robustness in our approach to imperfect parse ranking, I generalise the system to operate over n-best lists of MRSs. This setup yields greatly improved system performance and advances the state of the art on this task, i.e., makes my system retroactively the top performer in this specific competition.

The rest of this paper is organised as follows: in section 2, I describe the task of recognising textual entailment. I also briefly describe MRS representations, and mention previous work on RTE using MRS. In section 3, I analyse the shared task data, and implement an entailment decision component which takes as input MRS representations from the ERG. I then analyse the errors that the component makes. Finally, I compare my results to the actual winner of the 2010 shared task. In section 4, I generalise my approach to 10-best lists of MRSs.

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 76–84,
Gothenburg, Sweden, April 26-30 2014.

2 Background

In the following, I briefly review the task of recognising entailment between natural language sentences. I also show an example of an MRS representation, and mention some previous work on entailment recognition that has used MRSs.

2.1 Recognising Textual Entailment

Research on automated reasoning has always been a central topic in computer science, with much focus on logical approaches. Although there had been research on reasoning expressed in natural language, the PASCAL Recognising Textual Entailment (RTE) Challenge (Dagan et al., 2005) spurred wide interest in the problem. In the task proposed by the RTE Challenge, a system is required to recognise whether the meaning of one text can be inferred from the meaning of another text. Their definition of inference, or *textual entailment*, is based on the everyday reasoning abilities of humans rather than the logical properties of language.

The RTE Challenge evolved from the relatively simple task of making binary decisions about sentence pairs into more complex variants with many categories and multi-sentence texts. The data sets issued by the organisers over the years provide valuable research material. However, they contain a wide range of inference phenomena, and require both ontological and world knowledge. The data set that I have used for the present work, the PETE data set, focusses on syntactic phenomena, and does not require any knowledge about the state of the world or ontological relations.

2.2 Minimal Recursion Semantics

Minimal Recursion Semantics (MRS) (Copestake et al., 2005) is a framework for computational semantics which can be used for both parsing and generation. MRS representations are expressive, have a clear interface with syntax, and are suitable for processing. MRSs can be underspecified with regard to scope in order to allow a semantically ambiguous sentence to be represented with a single MRS that captures every reading. MRS is integrated with the HPSG English Resource Grammar (ERG) (Flickinger, 2000).

An MRS representation contains a multiset of relations, called *elementary predications* (EPs). An EP usually corresponds to a single lexeme, but can also represent general grammatical features. Each EP has a *predicate symbol* which, in the case of lexical predicates, encodes information about lemma, part-of-speech, and sense distinctions. An EP also has a *label* (also called *handle*) attached to it. Each EP contains a list of numbered arguments: `ARG0`, `ARG1`, etc. The value of an argument can be either a scopal variable (a handle which refers to another EP's label) or a non-scopal variable (events or states, or entities).

The `ARG0` position of the argument list has the EP's *distinguished variable* as its value. This variable denotes either an event or state, or a referential or abstract entity (e_i or x_i, respectively). Each non-quantifier EP has its unique distinguished variable.

Finally, an MRS has a set of *handle constraints* which describe how the scopal arguments of the EPs can be equated with EP labels. A constraint $h_i =_q h_j$ denotes equality modulo quantifier insertion. In addition to the indirect linking through handle constraints, EPs are directly linked by sharing the same variable as argument values. The resulting MRS forms a connected graph.

In figure 2, we see an MRS for the sentence *Somebody denies there are barriers* from the PETE development data (id 4116)[1]. The topmost relation of the MRS is `_deny_v_to`, which has two non-empty arguments: x_5 and h_{10}. x_5 is the distinguished variable of the relations `_some_q` and `person`, which represent the pronoun *somebody*. A handle constraint equates the sentential variable h_{10} with h_{11}, which is the label of `_be_v_there`. This last relation has x_{13} as its sole argument, which is the distinguished variable of `udef_q` and `_barrier_n_to`, the representation of *barriers*.

2.3 Previous Work on RTE using MRS

To my knowledge, MRS has not been used extensively in entailment decision systems. Notable examples of approaches that use MRSs are Wotzlaw and Coote (2013), and Bergmair (2010).

In Wotzlaw and Coote (2013), the authors present an entailment recognition system which combines high-coverage syntactic and semantic text analysis with logical inference supported by relevant background knowledge. Their system combines deep and shallow linguistic analysis, and transforms the results into scope-resolved

[1] The event and entity variables of the EPs often have grammatical features attached to them. I have removed these features from the MRS for the sake of readability.

$\langle\, h_1,$
h_4:proper_q$\langle 0:5\rangle$(ARG0 x_6, RSTR h_5, BODY h_7),
h_8:named$\langle 0:5\rangle$(ARG0 x_6, CARG *Japan*),
h_2:_deny_v_to$\langle 6:12\rangle$(ARG0 e_3, ARG1 x_6, ARG2 h_{10}, ARG3 i_9),
h_{11}:_be_v_there$\langle 19:22\rangle$(ARG0 e_{12}, ARG1 x_{13}),
h_{14}:udef_q$\langle 23:37\rangle$(ARG0 x_{13}, RSTR h_{15}, BODY h_{16}),
h_{17}:_real_a_1$\langle 23:27\rangle$(ARG0 e_{18}, ARG1 x_{13}),
h_{17}:_barrier_n_to$\langle 28:37\rangle$(ARG0 x_{13}, ARG1 i_{19})
$\{\, h_{15} =_q h_{17},\, h_{10} =_q h_{11},\, h_5 =_q h_8,\, h_1 =_q h_2 \,\}\,\rangle$

Figure 1: MRS for the sentence *Japan denies there are real barriers.*

$\langle\, h_1,$
h_4:person$\langle 0:8\rangle$(ARG0 x_5),
h_6:_some_q$\langle 0:8\rangle$(ARG0 x_5, RSTR h_7, BODY h_8),
h_2:_deny_v_to$\langle 9:15\rangle$(ARG0 e_3, ARG1 x_5, ARG2 h_{10}, ARG3 i_9),
h_{11}:_be_v_there$\langle 22:25\rangle$(ARG0 e_{12}, ARG1 x_{13}),
h_{14}:udef_q$\langle 26:35\rangle$(ARG0 x_{13}, RSTR h_{15}, BODY h_{16}),
h_{17}:_barrier_n_to$\langle 26:35\rangle$(ARG0 x_{13}, ARG1 i_{18})
$\{\, h_{15} =_q h_{17},\, h_{10} =_q h_{11},\, h_7 =_q h_4,\, h_1 =_q h_2 \,\}\,\rangle$

Figure 2: MRS for the sentence *Somebody denies there are barriers.*

MRS representations. The MRSs are in turn translated into another semantic representation format, which, enriched with background knowledge, forms the basis for logical inference.

In Bergmair (2010), we find a theory-driven approach to textual entailment that uses MRS as an intermediate format in constructing meaning representations. The approach is based on the assumptions that the syllogism is a good approximation of natural language reasoning, and that a many-valued logic provides a better model of natural language semantics than bivalent logics do. MRSs are used as a step in the translation of natural language sentences into logical formulae that are suitable for processing. Input sentences are parsed with the ERG, and the resulting MRSs are translated into ProtoForms, which are fully recursive meaning representations that are closely related to MRSs. These ProtoForms are then decomposed into syllogistic premises that can be processed by an inference engine.

3 Recognising Syntactic Entailment using MRSs

In this section, I briefly review the SemEval-2010 shared task that used entailment decision as a means of evaluating parsers. I then describe the entailment system I developed for the shared task data, and compare its results to the winner of the original task.

3.1 The PETE Shared Task

Parser Evaluation using Textual Entailments (PETE) was a shared task in the SemEval-2010 Evaluation Exercises on Semantic Evaluation (Yuret et al., 2010). The task involved building an entailment system that could decide entailment for sentence pairs based on the output of a parser. The organisers proposed the task as an alternative way of evaluating parsers. The parser evaluation method that currently dominates the field, PARSEVAL (Black et al., 1991), compares the phrase-structure bracketing of a parser's output with the gold annotation of a treebank. This makes the evaluation both formalism-dependent and vulnerable to inconsistencies in human annotations.

The PETE shared task proposes a different evaluation method. Instead of comparing parser output directly to a gold standard, one can evaluate *indirectly* by examining how well the parser output supports the task of entailment recognition. This strategy has several advantages: the evaluation is formalism-independent, it is easier for annotators to agree on entailment than on syntactic categories and bracketing, and the task targets semantically relevant phenomena in the parser output. The data are constructed so that syntactic analysis of the

sentences is sufficient to determine the entailment relationship. No background knowledge or reasoning ability is required to solve the task.

It is important to note that in the context of the PETE shared task, entailment decision is not a goal in itself, it is just a tool for parser evaluation.

The PETE organisers created two data sets for the task: a development set of 66 sentence pairs, and a test set of 301 pairs. The data sets were built by taking a selection of sentences that contain syntactic dependencies that are challenging for state-of-the-art parsers, and constructing short entailments that (in the case of positive entailment pairs) reflect these dependencies. The resulting sentence pairs were annotated with entailment judgements by untrained annotators, and only sentence pairs with a high degree of inter-annotator agreement were kept.

20 systems from 7 teams participated in the PETE task. The best scoring system was the Cambridge system (Rimell and Clark, 2010), with an accuracy of 72.4 %.

3.2 The System

My system consists of an entailment decision component that processes MRS representations as output by the ERG[2]. The entailment decision component is a Python implementation I developed after analysing the PETE development data.

The core idea is based on graph alignment, seeking to establish equivalence relations between components of MRS graphs. In a nutshell, if all nodes of the MRS corresponding to the hypothesis can be aligned with nodes of the MRS of the text, then we will call this relation MRS inclusion, and treat it as an indicator for entailment.[3] Furthermore, the PETE data set employs a limited range of "robust" generalisations in hypothesis strings, for example replacing complex noun phrases from the text by an underspecified pronoun like *somebody*. To accomodate such variation, my graph alignment procedure supports a number of "robust" equivalences, for example allowing an arbitrarily complex sub-graph to align with the graph fragment corresponding to expressions like *somebody*. These heuristic generalisations were designed in response to an in-depth analysis of the PETE development corpus, where I made the following observations for the sentences of positive entailment pairs (I use T_{sent} to mean the text sentence, and H_{sent} to mean the hypothesis sentence):

- H_{sent} is always shorter than T_{sent}.
- In some cases, H_{sent} is completely included in T_{sent}.
- Mostly, H_{sent} is a substructure of T_{sent} with minor changes:
  - T_{sent} is an active sentence, while H_{sent} is passive.
  - A noun phrase in T_{sent} has been replaced by *somebody*, *someone* or *something* in H_{sent}.
  - The whole of H_{sent} corresponds to a complex noun phrase in T_{sent}.

In addition, I noted that the determiner or definiteness of a noun phrase often changes from text to hypothesis without making any difference for the entailment. I also noted that, in accordance with the PETE design principles, the context provided by the text sentence does not influence the entailment relationship.

In the negative entailment pairs the hypothesis is usually a combination of elements from the text that does not match semantically with the text.

I examined treebanked MRS representations of the PETE development data in order to develop an entailment recognition heuristic. I found that by taking the EPs that have an *event variable* as their distinguished variable, I would capture the semantically most important relations in the sentence (the verbs). The heuristic picks out all EPs whose `ARG0` is an event variable from both the text and hypothesis MRSs—let us call them *event relations*. Then it tries to *match* all the event relations of the hypothesis to event relations in the text. In the following, T_{mrs} means the MRS for the text sentence, and H_{mrs} the MRS for the hypothesis. We say that two event relations match if:

1. they are the same or similar relations. Two event relations are the same or similar if they share the same predicate symbol, or if their predicate symbols contain the same lemma and part-of-speech.
2. and all their arguments match. Two arguments in the same argument position match if:

[2] I used the 1212 release of the ERG, in combination with the PET parser (Callmeier, 2000).

[3] On this view, bidirectional inclusion indicates that the two MRS graphs are isomorphic, i.e., logically equivalent.

- they are the same relation; or
- the argument in T_{mrs} represents a noun phrase and the argument in H_{mrs} is *somebody/someone/something*; or
- the argument in T_{mrs} is either a scopal relation or a conjunction relation, and the argument in the hypothesis is an argument of this relation; or
- the argument in H_{mrs} is not expressed.

Let us see how the heuristic works for the following sentence pair (PETE id 4116):

4116_T_{sent}: The U.S. wants the removal of what it perceives as barriers to investment; Japan denies there are real barriers.

4116_H_{sent}: Somebody denies there are barriers.

Figure 2 shows the MRS for 4116_H_{sent}. Figure 1 shows an MRS for the part of 4116_T_{sent} that entails 4116_H_{sent}: *Japan denies there are real barriers*. The heuristic picks out two relations in 4116_H_{mrs} that have an event variable as their distinguished variable: `_deny_v_to` and `_be_v_there`. It then tries to find a match for these relations in the set of event relations in 4116_T_{mrs}:

- The relation `_deny_v_to` also appears in 4116_T_{mrs}, and all its argument variables can be unified since their relations match according to the heuristic:
  - x_5 unifies with x_6, since `_some_q` and `person` (which represent *somebody*) match `proper_q` and `named` (which represent *Japan*[4])
  - h_{10} unifies with h_{10}, since they both (via the handle constraints) lead to the relation `_be_v_there`.
  - The variables i_9 and i_9 both represent unexpressed arguments, and so are trivially unified.
- The relation `_be_v_there` matches the corresponding relation in 4116_T_{mrs}, since their single argument x_{13} denotes the same relations: `udef_q` and `_barrier_n_to`.

[4]According to the heuristic, any proper name matches the pronoun *somebody*, so we do not have to consider the actual proper name involved.

This strategy enables us to capture all the core relations of the hypothesis. When examining the data one can see that, contrary to the design principles for the PETE data, some sentence pairs do require reasoning. The heuristic will fail to capture such pairs.

The ERG is a precision grammar and does not output analyses for sentences that are ungrammatical. Some of the sentences in the PETE data sets are arguably in a grammatical gray zone, and consequently the ERG will not give us MRS representations for such sentences. In some cases, errors in an MRS can also cause the MRS processing in the system to fail. Therefore, my system must have a fallback strategy for sentence pairs were MRSs are lacking or processing fails. The system answer NO in such cases, since it has no evidence for an entailment relationship.

For the development process I used both treebanked and 1-best MRSs.

3.3 Error analysis

Tables 1 and 2 show the entailment decision results for 1-best MRSs for the PETE development and test data. The ERG parsed 61 of the 66 pairs in the development set, and 285 of the 301 pairs in the test set. The five development set pairs that did not get a parse were all negative entailments pairs. Of the 16 test pairs that failed to parse, 10 were negative entailment pairs. The system's fallback strategy labels these as NO.

	gold YES: 38	gold NO: 28
sys YES	25	2
sys NO	13	26

Table 1: The results for 1-best MRSs for the PETE development data.

	gold YES: 156	gold NO: 145
sys YES	78	10
sys NO	78	135

Table 2: The results for 1-best MRSs for the PETE test data.

The implementation of the heuristic is fine-grained in its treatment of the transformations from text to hypothesis that I found in the PETE development sentences. Although I tried to anticipate possible variations in the test data set, it inevitably contained cases that were not covered by

the code. This meant that occasionally the system was not able to recognise an entailment.

However, most of the incorrect judgements were caused either by errors in the MRSs, or by features of the MRSs or the PETE sentence pairs that are outside the scope of my heuristic:

1. Recognising the entailment depends on information about coreferring expressions, which is not part of the MRS analyses.

2. The entailment (or non-entailment) relationship depends on something other than syntactic structure. Recognising the entailment requires background knowledge and reasoning. This means the entailment is really outside the stated scope of the PETE task.

3. For some of the PETE sentence pairs, the gold annotation can be discussed. The following pair (PETE id 2079) is labeled NO, but is structurally similar to sentence pairs in the data set that are labeled YES: *Also, traders are in better shape today than in 1987 to survive selling binges.* $\Rightarrow$ *Binges are survived.*

3.4 Results and Comparison to Shared Task Winner

At this point, we are ready to compare the results with the winner of the PETE shared task. Of the 20 systems that took part in the shared task, the best scoring participant was the Cambridge system, developed by Laura Rimell and Stephen Clark of the University of Cambridge (Rimell and Clark, 2010). Their system had an overall accuracy of 72.4 %. My focus here is on comparing the performance of the entailment systems, not the parsers.

The Cambridge system: The system consists of a parser and an entailment system. Rimell and Clark used the C&C parser, which can produce output in the form of grammatical relations, that is, labelled head-dependencies. They used the parser with the Stanford Dependency scheme (de Marneffe et al., 2006), which defines a hierarchy of 48 grammatical relations.

The Cambridge entailment system was based on the assumption that the hypothesis is a simplified version of the text. In order to decide entailment, one can then compare the grammatical relations—the SDs—of the two sentences[5]. If the SDs of the hypothesis are a subset of the SDs of the text, then the text entails the hypothesis. However, because the hypotheses in the PETE data are often not a direct substructure of the text, Rimell and Clark used heuristics to deal with alterations between sentences (in the following, I use T_{sd} and H_{sd} to mean the grammatical relations of text and hypothesis sentences, respectively):

1. If a SD in the hypothesis contains a token which is not in the text, this SD is ignored. This means that passive auxiliaries, pronouns, determiners and expletive subjects that are in H_{sd} but not in T_{sd} are ignored.

2. Passive subjects are equated with direct objects. This rule handles the PETE pairs where the active verb of the text has become a passive in the hypothesis.

3. When checking whether the SDs in H_{sd} are a subset of the SDs in T_{sd}, only subject and object relations are considered (core relations).

4. The intersection of SDs in T_{sd} and H_{sd} has to be non-empty (this is not restricted to subjects and objects).

To sum up: if core(H_{sd}) $\subseteq$ core(T_{sd}) and $H_{sd} \cap T_{sd} \neq \emptyset$, then T_{sent} entails H_{sent}.

Results for 1-best (automatically generated) test data: We can now compare the results from the system for 1-best test data with those of Cambridge.

In order to compare the test data results from my system with those of Rimell & Clark, I have to account for those sentence pairs that the ERG could not parse (16) and the MRS pairs that my system could not process (1). I use the same fallback strategy as Rimell & Clark, and let the entailment decision be NO for those sentence pairs the system cannot handle. For comparison, I also include the results for SCHWA (University of Sydney), the second highest scorer of the systems that participated in the shared task.

From the results in table 3 we can see that my system would have done well in the shared task. An accuracy of 70.7 % places the system a little

[5] In Rimell and Clark (2010), the authors used the abbreviation GR to mean the grammatical relations of the Stanford Dependency scheme. I use SD instead, to avoid confusion with the term GR as used by Carroll et al. (1999)

System	A	P	R	F1
Cambridge	72.4	79.6	62.8	70.2
My system	70.7	88.6	50.0	63.9
SCHWA	70.4	68.3	80.1	73.7

Table 3: The two top systems from the PETE shared task (Yuret et al., 2010) compared to my system. Accuracy (A) gives the percentage of correct answers for both YES and NO. Precision (P), recall (R) and F1 are calculated for YES.

ahead of SCHWA, the second best system. We also note that my system has a significantly higher precision on the YES judgements than the other two systems.

Resuls for gold/treebanked development data: In order to evaluate their entailment system, Rimell & Clark ran their system on manually annotated grammatical relations. Given a valid entailment decision approach and correct SDs, the system could in theory achieve 100 % accuracy. Cambridge achieved 90.9 % accuracy on these gold data. The authors noted that one incorrect decision was due to a PETE pair requiring coreference resolution, three errors were caused by certain transformations between text and hypothesis that were not covered by their heuristic, and two errors occured because the heuristic ignored some SDs that were crucial for recognising the entailments.

When I ran my system on treebanked MRSs for the PETE development data, it achieved an accuracy of 92.4 %, which is slightly better than the accuracy for Cambridge.

MRSs vs. grammatical relations: The information that the Cambridge system uses is word dependencies that are typed with grammatical relations. More specifically, Cambridge uses subject and object relations between words to decide entailment. Because the relations are explicit—we know exactly what type of grammatical relation that holds between two words—it is easy to select the relations in H_{sd} that one wants to check.

The EPs of MRSs are a mixture of lexical relations, and various syntactic and semantic relations. A lot of the grammatical information that is explicitly represented as SDs in the Stanford scheme is implicitly represented in MRS EPs as argument-value pairs. For example, the subject relation between *he* and the verb in *he runs* is represented as `(nsubj run he)` in Stanford notation. The corresponding representation in an MRS is `[ run_v_1 LBL: h ARG0: e ARG1: x ]`, where `ARG1` denotes the proto-agent of the verb. The assignment of semantic roles to arguments in EPs is not affected by passivisation or dative shift, whereas such transformations can cause differences in SDs. For sentence pairs where these phenomena occur, it is easier to match EPs and their arguments than the corresponding grammatical relations.

Cambridge heuristic vs. my heuristic: The Cambridge system checks whether the subject and object relations in H_{sd} also appear in T_{sd}. However, because their heuristic ignores tokens in the hypothesis that are not in the text, the system in certain cases does not check core relations that are crucial to the entailment relationship.

My system checks whether the event relations in H_{mrs} also appear in T_{mrs}, and whether their arguments can be matched. Whereas the Cambridge system ignores tokens in the hypothesis that have no match in the text, my heuristic has explicit rules for matching arguments that are different. It makes my system more vulnerable to unseen cases, but at the same time makes the positive entailment decisions more well-founded. It leads my system to make fewer mistakes on the NO entailments than both the Cambridge system and SCHWA.

In their paper, Rimell & Clark do not provide an error analysis for the PETE test set, so I cannot do a comparative error analysis with my system. However, they go into detail on some analyses and mention some errors that the system made on the development data (both automatically generated and gold-standard), and I can compare these to my own results on the development data. (I will only look at those analyses where there are significant differences between Cambridge and my system.)

PETE id 5019: *He would wake up in the middle of the night and fret about it.* $\Rightarrow$ *He would wake up.* The Cambridge system recognises this correctly, but the decision is based only on the single SD match `(nsubj would he)`. The other SDs are ignored, since they are non-core according to the heuristic. In my system, the YES decision is based on matching of both the relation `_would_v_modal` which has `_wake_v_up` as its scopal argument, and `_wake_v_up` itself with its

pronoun argument.

PETE id 3081.N: *Occasionally, the children find steamed, whole-wheat grains for cereal which they call "buckshot". ⇒ Grains are steamed.* The transformation of *steamed* from an adjective in T_{sent} to a passive in H_{sent} was not accounted for in the Cambridge heuristic, and the system failed to recognise the entailment. In the MRS analyses for these sentences, *steamed* gets exactly the same representation, and my entailment system can easily match the two.

The Cambridge paper mentions that two of the errors the entailment system made were due to the fact that a non-core relation or a pronoun in the hypothesis, which Cambridge ignores, was crucial for recognising an entailment. The paper does not mention which sentences these were, but it seems likely that they would not pose a problem to my system.

4 Using 10-best MRSs

So far, I have used only one MRS per sentence in the entailment decision process. The entailment decisions were based on the best MRSs for a sentence pair, either chosen manually (treebanked MRSs) or automatically (1-best MRSs). In both cases, it can happen that the MRS chosen for a sentence is not actually the best interpretation, either because of human error during treebanking, or because the best MRS is not ranked as number one.

I also noticed that many of the incorrect decisions that the system made were caused either by errors in the MRSs or by incompatible analyses for a sentence pair. In both cases, the correct or compatible MRS could possibly be found further down the list of analyses produced by the ERG. These shortcomings can perhaps be remedied by examining more MRS analyses for each sentence in a pair.

When doing n-best parsing on the PETE data sets, we can expect a high number of analyses for the text sentences, and fewer analyses for the shorter hypotheses. By setting n to 10, I hope to capture a sufficient number of the best analyses. With 10-best parsing, I get on average 9 analyses for the text sentences, and 3 analyses for the hypotheses.

I use a simple strategy for checking entailment between a set of MRSs for the text and a set of MRSs for the hypothesis: If I can find one case of entailment between two MRSs, then I conclude that the text entails the hypothesis.

In table 4, I compare my previous results with those that I get with 10-best MRSs. As we can see, the system manages to recognise a higher number of positive entailment pairs, but the precision goes down a little. Using 10-best MRSs ensures that we do not miss out on positive entailment pairs where an incorrect MRS is ranked as number one. However, it also increases the number of spurious entailments caused by MRSs whose event relations accidentally match. Variation of *n* allows trading off precision and recall, and *n* can possibly be tuned separately for texts and hypotheses.

When we compare 10-best entailment checking to the PETE shared task results, we see that my results improve substantially over the previously highest reported performance. My system scores about 4 accuracy points higher than the system of Rimell & Clark, and more than 5 points for F1.

System	A	P	R	F1
One MRS	70.7	88.6	50.0	63.9
10-best	76.4	81.4	70.5	75.5

Table 4: Here I compare system results for one MRS and 10-best MRSs. Accuracy (A) gives the percentage of correct answers for both YES and NO. Precision (P), recall (R) and F1 are calculated for YES.

5 Conclusions and Future Work

In this paper, I have demonstrated how to build an entailment system from MRS graph alignment, combined with heuristic "robust" generalisations. I compared my results to the winner of the 2010 PETE shared task, the Cambridge system, which used grammatical relations as the basis for entailment decision. I performed an in-depth comparison of types and structure of information relevant to entailment in syntactic dependencies vs. logical-form meaning representations. The system achieved competitive results to the state of the art. Results on gold-standard parser output suggests substantially better performance in my entailment system than the PETE shared task winner.

I also generalised the approach to using n-best lists of parser outputs. Using 1-best output makes entailment decision vulnerable to incorrect MRS analyses being ranked as number one. Using n-best can counterbalance this prob-

lem. With 10-best MRSs, a significant improvement was achieved in the performance of the entailment decision system. The n-best setup offers the flexibility of trading off precision and recall.

With the 10-best MRS lists, I used a simple strategy for entailment decision: if one MRS pair supports a YES decision, we say that we have entailment. It would be interesting to explore more complex strategies, such as testing all the MRS combinations for a sentence pair for a certain n, and decide for the majority vote. One could also make use of the conditional probabilities on parser outputs, for instance by multiplying the probabilities for each MRS pair, summing up for YES vs. NO decisions, and setting a threshold for the final decision.

Acknowledgments

I am grateful to my supervisors Jan Tore Lønning and Stephan Oepen for suggesting this task, and for their valuable advice on my work. I also appreciate the thorough comments made by the three anonymous reviewers.

References

Richard Bergmair. 2010. *Monte Carlo Semantics: Robust Inference and Logical Pattern Processing with Natural Language Text*. Ph.D. thesis, University of Cambridge.

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Speech and natural language: proceedings of a workshop, held at Pacific Grove, California, February 19-22, 1991*, page 306. Morgan Kaufman Pub.

Ulrich Callmeier. 2000. PET. A platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering*, 6(1):99108, March.

John A. Carroll, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC)*.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation*, 3(2):281–332.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, Genoa, Italy.

Dan Flickinger. 2000. On building a more effcient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Laura Rimell and Stephen Clark. 2010. Cambridge: Parser Evaluation using Textual Entailment by Grammatical Relation Comparison. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*.

Andreas Wotzlaw and Ravi Coote. 2013. A Logic-based Approach for Recognizing Textual Entailment Supported by Ontological Background Knowledge. *CoRR*, abs/1310.4938.

Deniz Yuret, Aydin Han, and Zehra Turgut. 2010. SemEval-2010 Task 12: Parser Evaluation using Textual Entailments. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 51–56. Association for Computational Linguistics.